# Provider Competition in a Dynamic Setting

**by Marie ALLARD, Pierre Thomas LÉGER and Lise ROCHAIX**

**Cahier de recherche nº IEA-04-07**
August 2004

ISSN : 0825-8643

---


*La publication de ce Cahier de recherche a été rendue possible grâce à des subventions d'aide à la publication et à la diffusion de la recherche provenant des fonds de HEC Montréal.*
*Direction de la recherche, HEC Montréal, 3000, chemin de la Côte-Sainte-Catherine, Montréal (Québec) Canada H3T 2A7.*

# Provider Competition in a Dynamic Setting*

Marie Allard† Pierre Thomas Léger‡ Lise Rochaix§

August 5 2004

**Abstract**

In this paper, we examine provider and patient behaviour where effort is non-contractible and where competition between providers is modeled in an explicit way. More specifically, we construct a model where physicians repeatedly compete for patients and where patients' outside options are solved for in equilibrium. In our model, physicians are characterized by an individual-specific ethical constraint which allows for unobserved heterogeneity in the physicians market. By doing so, we introduce uncertainty in the patient's likely treatment if he were in fact to leave his current physician to seek care elsewhere. We find that competition between providers may serve as an important incentive for physicians in treating their patients with desired levels of care.

JEL classification: I10, I18, J24, C30

Keywords: Physician Payment Mechanisms, Physician heterogeneity, Competition, Information Asymmetry, Insurance.

*We thank seminar participants at the 4th Health Economics Workshop (Oslo), DELTA/CREST-LEI (Paris), 3e Journées d'économie publique LAGV (IDEP-Marseille) and HEC Montréal. Allard thanks the CEPREMAP, where part of this work was undertaken, for its hospitality. The usual caveats apply.

†HEC Montreal

‡HEC Montreal, CIRANO and CIRPÉE; corresponding author: pierre-thomas.leger@hec.ca

§IDEP-GREQAM-Université de la Méditerranée (Aix-Marseille II)

# 1 Introduction

The provision of medical services includes several forms of care of which some are unobservable and thus non-contractible. Several studies have examined different mechanisms, such as physician monitoring and/or payment schemes, which seek to encourage the efficient provision of these unobservable forms of care such as physician time and effort. In this paper, we examine the role of competition as an alternative way of dealing with this issue. More specifically, we construct a model where physicians repeatedly compete for patients and where patients' outside options are solved for in equilibrium. In our model, each physician is characterized by an individual-specific ethical constraint which specifies the minimal amount of effort to be provided. These ethical constraints allow for unobserved heterogeneity in the physicians market. By doing so, we introduce uncertainty in the patient's likely treatment if he were in fact to leave his current physician to seek care elsewhere. We find that competition between providers may serve as an important incentive for physicians in treating their patients with desired levels of care.

Although patients, doctors and insurers may be able to observe certain components of care (such as hospitalizations, testing, pharmaceuticals...), several valued forms of care are unobservable by third-parties and thus non-contractible. The presence of competition and certain types of payment schemes may, however, serve as important mechanisms to encourage desired provision of such non-contractible care, in a setting characterized by information asymmetry. In our model, we build on several papers (Ellis and McGuire, 1986; Ma, 1994; Ma and McGuire, 1997; Ellis, 1998; Gal-Or, 1999) while exploiting competition between similar providers (for example, between GPs) in a specific way. According to Gaynor and Vogt (2000), competition has been somewhat ignored in the literature partly because of the lack of concentration in the physicians market, i.e., the market is unlikely to exhibit anti-competitive behaviour. However, the presence of information asymmetry between patients and physicians, the proliferation of prospective payments which may encourage sub-optimal care, and the discretionary powers held by physicians, nonetheless point to a role for

competition and/or monitoring in the physicians market. Although monitoring (either directly or through medical malpractice litigation) may be a way to address these market imperfections, there is still a need to study other mechanisms such as competition in order to determine how to achieve efficient provision of care.[1]

Our work is related to several papers on competition in the physicians market. In Rochaix (1989), the patient's ability to consult a competing physician imposes an implicit constraint on his physician's discretionary power. More specifically, physicians risk losing their patients if their diagnosis differs greatly from their patients' prior expectations about illness severity. The threat of losing patients leads physicians to recommend a treatment intensity that is closer to the full information solution (a result which holds in the presence of only a small number of informed patients). Rochaix, however, does not deal with the issue of non-observable (and thus, non-contractible) effort. In Allard *et al.* (2001), the authors study compensation of health-care providers in a principal-agent framework where information asymmetry exists between providers and the regulatory agent. In this model, physicians are differentiated by their productivity. Patients, who are assumed to be identical, choose the physician who offers them the greatest net benefit. In equilibrium, competition in the physicians market equalizes net benefits among patients, i.e., the 'market constraint' leads physicians to exert non-contractible effort in order to attract patients. Our paper differs from this one in several respects, most notably, by introducing patient heterogeneity. Furthermore, unlike Allard *et al.,* our model can generate both treatment heterogeneity and patient turnover in equilibrium. Finally, Ma and McGuire (1997) examine the role of competition by having physicians compete with an exogenously given outside option (i.e., where the patient can obtain a given utility if he decided to leave) and by introducing patient heterogeneity with respect to their out-of-pocket cost for using different physicians. Endogenizing this outside option is at the heart of our model.

In this paper, we find that under certain conditions competition may lead physicians to treat their patients with desired levels of care independently of their type (i.e., independently of their

[1] For a discussion of monitoring see Léger (2000). For a discussion of medical malpractice see Danzon (2000).

ethical constraint) - thus leading to stable physician-patient relationships. In the presence of non-trivial switching costs, however, the effect of competition is somewhat dampened. In such a case, while certain patients will receive more care than others (i.e., the equilibrium will be characterized by heterogeneity in treatments), stable physician-patient relationships still exist.[2] Competition will, nonetheless, lead to a lower-bound in effort provided, where a mass of physicians will provide effort beyond what is determined by their ethical constraint. Finally, under certain conditions such as excess demand in the physicians market or relatively myopic physicians, the equilibrium may be characterized by heterogeneity in treatments as well as some unstable physician-patient relationships. Thus, under certain conditions, some patient turnover will occur in equilibrium. Even if the presence of such an equilibrium, competition will nonetheless induce a mass of physicians to treat their patients beyond the level of care determined by their ethical constraint.

The remainder of the paper is organized as follows. In section 2, we describe the model. In section 3, we solve the model in a static setting. We resolve the model in a repeated-game setting in section 4. Conclusions are drawn in section 5.

## 2 The Model

In this section we introduce a dynamic model characterizing the relationship between physicians, patients and insurance providers. As in Ma and McGuire (1997), treatment following an illness requires two forms of medical input: (i) observable medical care $q$, and (ii) unobservable physician effort $\epsilon$. Medical care $q$ is defined as any form of observable and contractible medical treatment. On the other hand, effort $\epsilon$ may be thought of as all valued forms of care which are not observable to third parties and thus non-contractible. These forms of care may include the physician's time

[2] According to Ellis and McGuire (1986): 'Available evidence suggests that health care consumers do a very limited amount of shopping around among physicians, and that, having chosen a physician, consumers accept most physician's recommendations quite passively.'(p.144). If one assumes that competition between providers necessarily translates itself into patients shopping around among providers, then the aforementioned evidence suggests little competition in the market for physician services. In our model, however, it is not shopping around but rather the threat of moving from one physician to another that creates competitive pressures between providers. Thus, as in our results where the equilibria are characterized by stable patient-physician relationships, the physicians market may be very competitive without ever exhibiting patient shopping.

and effort spent in researching and providing the appropriate treatment, monitoring the patient's progress and communicating with the patient (see Wedig *et al.*, 1989). We further assume a mixed physician payment scheme which consists of both a per-unit-of-$q$ reimbursement and a prospective payment. This prospective component will ultimately serve to compensate physicians for the effort they exert, given that this form of care cannot be reimbursed on a per-unit basis.

Before competition (for patients) begins, a population of measure one of patients is assumed to be equally allocated to a population of measure one of physicians. Competition is introduced in our model by adopting a multi-period setting where patients can move from one physician to another. Because we adopt such a framework, our model is best suited to potentially long-term relationships between patients and providers (for example, between patients and their family practitioners or, in the case of a chronic illness, between patients and their specialists).

The timing of the game is as follows:

Stage 1:

The physician-payment and insurance parameters are contracted upon. It is at this stage that the patient purchases an actuarially-fair insurance policy at a premium $\alpha$.

Stage 2:

With probability $\pi$, the patient becomes ill and requires medical treatment. If the patient is ill, he draws $\theta$ from a known distribution of illness $F(\theta)$.[3] We assume that the patient perfectly observes his illness severity which is not observable to the third-party payer. If the patient is not ill, the 'period' ends (i.e., the patient does not seek medical treatment, remains healthy for one full period and returns, in the repeated-game setting, to stage 1 in the next period).

Stage 3:

A patient with illness severity $\theta$ seeks medical treatment. In our model it is assumed that $\epsilon$ and $q$ are chosen simultaneously by the physician and the patient, respectively, i.e., neither patient

[3] In this setup, we can think of $\theta$ as representing a single illness with a severity distribution or a composite measure which maps different types of illnesses and their severity into a single dimension.

nor physician can base his or her decision on the other's choice.[4,5] We assume, however, that the quantity $q$ is purchased (on behalf of the patient) by the physician at a cost of $\omega$ per unit.

Stage 4:

Once medical care and effort have been provided, the patient's ex post health $H$ (given by the health production function $h(\theta, q, \epsilon)$) is revealed. We assume that ex post health is perfectly observable to the patient yet unobservable to the third party. Once the physician has treated the patient: (i) the patient pays $\gamma pq$ where $\gamma$ denotes the co-payment rate and $p$ denotes the price per-unit of quantity $q$, and (ii) the physician receives a net payment $(p - \omega)$ for each unit of quantity $q$ provided and a prospective payment $\delta$ which serves to compensate for effort.[6]

Stage 5:

Because the patient observes his illness severity $\theta$, the quantity of medical care $q$ provided and his health outcome $H$, he can infer his physician's effort $\epsilon$. Based on this information, the patient may choose to leave his current physician. For simplicity, we assume that each period is characterized by a new draw from the illness distribution, i.e., we exclude the 'dynamic' aspect of health.[7]

We next describe each player in greater detail.

The Patient:

The patient per-period expected utility is given by:

$$EU = (1 - \pi)U(C, H^0) + \pi \int_\theta U(C, h(\theta, q, \epsilon))dF(\theta), \tag{1}$$

where

$$C = I - \alpha - \gamma pq. \tag{2}$$

[4] We differ from Ma and McGuire (1997) in this respect, i.e., we relax their somewhat restrictive assumption that the patient observes the effort provided by his physician before choosing the quantity of medical care.

[5] Allowing the patient to choose the quantity $q$ is equivalent to the physician proposing a schedule of treatments and prices. Because greater levels of $q$ are associated with greater costs (i.e., a higher co-payment), the patient will choose the quantity which maximizes his expected utility.

[6] Thus, in this framework, physicians only receive payment if the patient is ill and seeks medical care.

[7] Because the patient draws from the illness severity distribution independently in each period, cream-skimming issues are not dealt with here. That is, because all patients are identical before each period begins, physicians will not be able to select less or more costly patients.

We assume a separable utility function for $U(C, H)$ :

$$U(C, H) = u(c) + h(\theta, q, \epsilon),$$

where $u' > 0$ and $u'' < 0$. Furthermore, in (2) $C$ denotes the patient's consumption while $I$ denotes the state-independent income. We define $H^0 \equiv h(0, 0, 0)$ to be the patient's health in the absence of illness.

The Physician:

In our model, we introduce unobserved heterogeneity in the physicians market in a simple way. Each physician is characterized by a $\lambda$ parameter where $\lambda \in [0, 1]$. If for a given illness severity $\theta$, the patient were to choose in stage 3 an effort $\widetilde{\epsilon}$ (henceforth referred to as the patient's desired level of effort), a physician $\lambda$ would never be willing to provide less than $\lambda\widetilde{\epsilon}(\theta)$.[8] For example, a physician with $\lambda = 1$ would never be willing to provide less than the patient's desired level of effort ($\epsilon = \widetilde{\epsilon}(\theta)$). However, a physician with $\lambda = 0$ could provide the minimal amount of effort possible ($\epsilon = 0$).[9] Thus, each physician will be characterized by an ethical constraint which gives the minimum proportion of the desired effort level to be provided. We also assume that physician types are distributed according to a known distribution $\Gamma(\lambda)$.

Each physician is assumed to have a per-patient per-period utility $V$ which is increasing in income $M$ and decreasing in effort $\epsilon$. Thus, the physician's per-patient per-period expected utility is given by:

$$EV = (1 - \pi)V(0, 0) + \pi V(M, \epsilon), \tag{3}$$

where $M = \delta + (p - \omega)q$ when the patient seeks medical treatment. We assume a separable utility function for $V(M, \epsilon)$:

$$V(M, \epsilon) = M - c(\epsilon),$$

[8] It is important to note that for every co-payment rate $\gamma$ and every illness severity $\theta$, there exists a patient's utility maximizing $q$ and $\epsilon$ in Stage 3. Because the desired effort level $\widetilde{\epsilon}$ is dependent on the co-payment rate, it cannot be thought of as some medically-justified amount of effort.

[9] One can think of effort $\epsilon = 0$ as the minimal amount of effort below which the physician's effort would be observably insufficient.

where $c' > 0$ and $c'' > 0$.

The Insurer:

We assume that the market for insurance is perfectly competitive. The actuarially-fair health-insurance premium for physician services is thus given by:

$$\alpha = \pi \int_{\theta} ((1-\gamma)pq(\theta) + \delta(\epsilon(\theta)))dF(\theta). \tag{4}$$

where $q(\theta)$ and $\epsilon(\theta)$ denote the quantity of medical services and effort in equilibrium.

## 3 The Static Framework

In this section, we examine the static setting by shutting down Stage 5 in the game described above. Examining our model without its competitive feature will serve as a benchmark.

It is well known in the literature that the first-best health insurance policy would provide state-contingent treatments (in our case, illness contingent levels of $q$ and $\epsilon$). In our case, optimal illness-contingent levels of $q$ and $\epsilon$ can be obtained by solving the patient's ex ante problem. That is, optimal levels of $q$ and $\epsilon$ can be obtained by maximizing the patient's expected utility (1) subject to his budget constraint (2), the physician-participation constraint (that will be satisfied if the physician's expected utility (3) is greater than some exogenously given value $\overline{V}$), and an actuarially-fair health-insurance premium (4). However, a state-contingent contract of this type is infeasible given that illness severity, effort levels and post-treatment health are not verifiable and thus non-contractible (Arrow, 1963).

As noted above, the patient observes his illness severity and his ex post health but does not observe his physician's type. Also recall that the physician chooses effort level $\epsilon$ while the patient simultaneously chooses medical care $q$. It is obvious that in a static setting the physician will never wish to provide effort beyond the minimum amount determined by her ethical constraint, i.e., for a given illness severity $\theta$, the physician $\lambda$ will provide $\lambda\widetilde{\epsilon}(\theta)$ irrespective of the prospective

payment. This is simply because increasing the effort beyond the minimum amount, which is utility decreasing for the physician, does not yield a larger prospective payment for the physician.

For a given co-payment $\gamma$ and a specific realization of $\theta$, the patient's expectation with respect to his physician's effort is given by $E_\lambda(\lambda\widetilde{\epsilon}(\theta)) = \widehat{\lambda}\widetilde{\epsilon}(\theta)$ where $\widehat{\lambda} = \int_0^1 \lambda d\Gamma(\lambda)$. Thus a patient with illness $\theta$ solves:

$$\max_q U(I - \alpha - \gamma pq, h(\theta, q, \widehat{\lambda}\widetilde{\epsilon}(\theta))). \tag{5}$$

For a given co-payment $\gamma$ and a specific illness severity $\theta$, the equilibrium will be characterized by homogeneity in quantities $q^*(\theta)$ chosen by the patients yet, heterogeneity in efforts $\epsilon^*(\theta)$ provided by the physicians (where the equilibrium efforts will be distributed between 0 and $\widetilde{\epsilon}(\theta)$). As a result, how much effort the patient receives is simply a function of his illness severity and the physician type he has been assigned to.

To ensure the participation of all physicians (i.e., irrespective of type), the prospective payment must (at least) compensate effort provided by the physician of type $\lambda = 1$, i.e. $\delta(\theta) \geq c(\widetilde{\epsilon}(\theta))$. We henceforth set $\delta(\theta) = c(\widetilde{\epsilon}(\theta))$. If $\theta$ were observable, an illness-specific prospective payment $\delta(\theta)$ would have to be paid to all physicians irrespective of their type. However, given that the illness severity is not observable by the insurer, the equilibrium prospective payment $\delta^*$, which is paid to the physician prior to the realization of $\theta$, must be illness independent and based on its expectation, i.e.,

$$\delta^* = \int_\theta \delta^*(\theta) dF(\theta) = \int_\theta c(\widetilde{\epsilon}(\theta)) dF(\theta). \tag{6}$$

Next, the actuarially-fair insurance premium $\alpha$ is given by:

$$\alpha(\gamma) = \pi \int_\theta ((1-\gamma)pq^*(\theta)) dF(\theta) + \pi\delta^*. \tag{7}$$

Given our assumption of perfect competition in the insurance market, insurers will be indifferent between all co-payment levels (i.e., each co-payment level is associated with an insurance premium that yields zero expected profits). As a result, the equilibrium co-payment $\gamma^*$ will maximize the patient's expected utility. This $\gamma^*$ balances the expected utility gains of more complete insurance

with the utility loss of a higher insurance premium. Thus, the equilibrium actuarially-fair insurance premium $\alpha^*$ is simply given by (7) evaluated at $\gamma^*$.

Given the results provided above, we can characterize both the patients' and physicians' ex post utility. The patient's ex post utility is given by:

$$U(I - \alpha^* - \gamma^* p q^*(\theta), h(\theta, q^*(\theta), \lambda\widetilde{\epsilon}(\theta))), \tag{8}$$

where we recall that the quantity $q^*$ is chosen based on the realization of $\theta$ and the *expected* effort level to be provided by his physician. Ex post health, however, is a function of the realization of illness severity $\theta$, $q^*$ and the *true* effort provided by the physician. Thus, if the patient's physician is of a type greater than the expected type ($\lambda > \widehat{\lambda}$), then the patient will be treated with more effort than expected. In such a case, the patient will have chosen a quantity $q^*$ which is too large (small) if $q$ and $\epsilon$ are substitutes (complements).

In this setting all physicians receive the same compensation (i.e., irrespective of their type): $M = \delta^* + (p - \omega)q^*(\theta) = \delta^*$ if $p = \omega$. However, physician ex post utility is type dependent, i.e.,

$$V(\delta^* + (p - \omega)q^*, \lambda\widetilde{\epsilon}(\theta)) = \delta^* - c(\lambda\widetilde{\epsilon}(\theta)). \tag{9}$$

Thus, in equilibrium, all but the physician with $\lambda = 1$ will receive a prospective payment which over-compensates for effort provided (in expected terms).

The above result, where all physicians provide their respective minimum effort, is consistent with Ma and McGuire's statement that: 'the alternative assumptions - that physician effort decision is made either simultaneously with, or after the patient's quantity decision- are unpalatable: in both cases, neither the patient's quantity choice nor the payment contract can provide any incentive for the physician to undertake costly actions.'(p. 690). In the next section we show that this is not necessarily the case when competition is introduced in a dynamic setting. That is, we show that physicians may undertake costly effort even if physician effort is chosen simultaneously with the patient's quantity decision when they repeatedly compete for patients.

# 4 The Dynamic Framework

In this section, we turn our attention to a richer model where competition between providers plays a central role. In a repeated game setting, the patient's ability to move from one physician to another may serve to encourage physicians to provide treatment levels beyond those determined by their ethical constraints.

Although many equilibria may be supported by non-credible threats, such equilibria are of little interest in a dynamic setting. Take, for example, a patient who follows a variant of the trigger strategy which states that he will leave his current physician if he is not treated with desired levels of care, i.e., if his ex post utility is less than $U(I-\alpha-\gamma p\widetilde{q}(\theta), h(\theta, \widetilde{q}(\theta), \widetilde{\epsilon}(\theta)))$, $\forall\theta$. This threat of leaving is only credible if the patient is indeed willing to leave in the presence of 'sub-desired' care.[10] Given the patient's strategy, the physician should provide the desired effort level under the condition that the discounted expected utility of providing desired effort is greater than the discounted expected utility of providing minimal effort and losing the patient. Although, under certain conditions, such a trigger strategy may yield an equilibrium, it may be supported by a threat which is not credible. In order to determine whether a patient's threat is in fact credible, we derive what the patient could expect to receive if he did in fact leave for a competing physician. We next compare this with what he could expect to receive if he were to remain with his current one. By doing so, we limit ourselves to examining equilibria which are supported by credible threats.

## 4.1 The Patients' and Physicians' Strategies

In this section, we define the patients' and physicians' strategies in a repeated-game framework.

**The Patients' Strategy:**

If the patient left his current physician, he would receive at the end of the first period:

$$U(I-\alpha-\gamma pq^{*}-\kappa, h(\theta, q^{*}, \epsilon)), \tag{10}$$

[10] Given that the patient observes his illness severity and decides on $q$, the above strategy is equivalent to a strategy based on ex post health.

and expect to receive in the future (at least):

$$\sum_{t=2}^{\infty} \rho^{t-1} U^{Leave} = \sum_{t=2}^{\infty} \rho^{t-1} \int_{\theta} U(I - \alpha - \gamma p q^*, h(\theta, q^*, \epsilon^{\exp}(\theta))) dF(\theta), \tag{11}$$

where $\kappa$ is included to represent financial and/or psychic costs associated with moving from one physician to another, and where $\rho$ denotes the patient's discount factor. In (11), $\epsilon^{\exp}(\theta)$ denotes the patient's expectation about the effort to be provided by the outside physician if he were to leave his current one. More specifically, $\epsilon^{\exp}(\theta) = \int_{\lambda} \epsilon_{\lambda}(\theta) d\Gamma(\lambda)$, where $\epsilon_{\lambda}(\theta)$ is the effort that physician $\lambda$ will provide in equilibrium. It is important to note that $q^*$ in (10) is based both on the current period illness severity and the expected effort provided by the patient's current physician. However, $q^*$ in (11) is based both on the illness severity and on the expected effort provided by the outside physician ($\epsilon^{\exp}(\theta)$).

In order to characterize the present value of not leaving, we must define how patients form their expectations regarding future effort levels to be provided by their current physician. Indeed, recall that the patient observes $\theta$, chooses $q$, observes ex post health and thus can infer the effort provided to him by his physician in the current period. Although a patient can not perfectly infer his physician's type, he can infer to some extent what type his physician is not. That is, a patient who draws $\theta$ can always infer an upper bound for his physician's type. As a result, a physician who provides $\epsilon$ (given $\theta$) must be identified by a $\lambda \in [0, \epsilon/\widetilde{\epsilon}]$ where $\widetilde{\epsilon}$ is the desired effort level for the particular value of $\theta$. We denote $\lambda^{\max} = \epsilon/\widetilde{\epsilon}$. In the following sections, patients will base their expectations regarding their current physician's future behaviour on this $\lambda^{\max}$.[11] While basing future behaviour on $\lambda^{\max}$, rather than any other value in the interval $[0, \lambda^{\max}]$, may appear somewhat limiting and arbitrary, we show later on that these are the only expectations which survive in equilibrium under reasonable assumptions.

As a result, if the patient remained with his current physician, he would receive in the current period:

[11] Although it is possible for a physician for whom $\lambda^{\max}\widetilde{\epsilon}(\theta) < \epsilon^{\exp}(\theta)$ to provide, in the future, effort greater than $\epsilon^{\exp}(\theta)$, those for whom $\lambda\widetilde{\epsilon}(\theta) > \epsilon^{\exp}(\theta)$ have no choice but to do so. Therefore, it is reasonable to believe that, *ceteris paribus*, the latter will provide greater effort in the future than the former.

$$U(I-\alpha-\gamma pq^*, h(\theta, q^*, \epsilon)), \tag{12}$$

and expect to receive in the future (at least):

$$\sum_{t=2}^{\infty}\rho^{t-1}U^{Stay} = \sum_{t=2}^{\infty}\rho^{t-1}\int_{\theta}U(I-\alpha-\gamma pq^*, h(\theta, q^*, \lambda^{\max}\widetilde{\epsilon}(\theta)))dF(\theta). \tag{13}$$

We now write the patient's strategy based on (10), (11), (12), and (13). That is, the patient will be willing to leave his current physician if:

$$\begin{aligned} & U(I-\alpha-\gamma pq^*-\kappa, h(\theta, q^*, \epsilon)) + \sum_{t=2}^{\infty}\rho^{t-1}\int_{\theta}U(I-\alpha-\gamma pq^*, h(\theta, q^*, \epsilon^{\exp}(\theta)))dF(\theta) \\ > \;& U(I-\alpha-\gamma pq^*, h(\theta, q^*, \epsilon)) + \sum_{t=2}^{\infty}\rho^{t-1}\int_{\theta}U(I-\alpha-\gamma pq^*, h(\theta, q^*, \lambda^{\max}\widetilde{\epsilon}(\theta)))dF(\theta). \end{aligned} \tag{14}$$

If we assume, for the time being, that transaction costs are arbitrarily small (i.e., $\kappa = 0$), we can rewrite (14) as:

$$\begin{aligned} & \sum_{t=2}^{\infty}\rho^{t-1}\int_{\theta}U(I-\alpha-\gamma pq^*, h(\theta, q^*, \epsilon^{\exp}(\theta)))dF(\theta) \\ > \;& \sum_{t=2}^{\infty}\rho^{t-1}\int_{\theta}U(I-\alpha-\gamma pq^*, h(\theta, q^*, \lambda^{\max}\widetilde{\epsilon}(\theta)))dF(\theta). \end{aligned} \tag{15}$$

That is, the patient will leave (stay with) his current physician if $\epsilon^{\exp}(\theta) > (\leq)\ \lambda^{\max}\widetilde{\epsilon}(\theta)$.

**The Physicians' Strategy:**

We now turn our attention to the physician's strategy. A physician for whom $\lambda\widetilde{\epsilon}(\theta) \geq \epsilon^{\exp}(\theta)$ will provide effort according to her ethical constraint (i.e., $\lambda\widetilde{\epsilon}(\theta)$). By doing so, the physician will be minimizing her effort costs and will not lose her patient. However, a physician for whom $\lambda\widetilde{\epsilon}(\theta) < \epsilon^{\exp}(\theta)$ (i.e., for whom the effort determined by her ethical constraint is less than the effort the patient could expect if he left for an outside physician) will provide $\epsilon^{\exp}(\theta)$ if providing such effort yields greater discounted expected utility than providing her minimum effort and losing her patient i.e., if:

$$V(\delta, \epsilon^{\exp}(\theta)) + \sum_{t=2}^{\infty}\beta^{t-1}\int_{\theta}V(\delta, \epsilon^{\exp}(\theta))dF(\theta) \geq V(\delta, \lambda\widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty}\beta^{t-1}V_t^{DEV}. \tag{16}$$

In (16) $\beta$ denotes the physician's discount rate. Furthermore, we define $\sum_{t=2}^{\infty}\beta^{t-1}V_t^{DEV}$ as the future discounted expected utility associated with losing one's patient.[12] Below, we rule out the two extreme cases where either (i) the physician does not find a replacement for the lost patient in the long run (for example, because of high excess supply) (i.e.,$\sum_{t=2}^{\infty}\beta^{t-1}V_t^{DEV}=\sum_{t=2}^{\infty}\beta^{t-1}V(0,0)$) , or (ii) the physician finds an immediate replacement for the lost patient (for example, because of high excess demand) (i.e., $\sum_{t=2}^{\infty}\beta^{t-1}V_t^{DEV}=\sum_{t=2}^{\infty}\beta^{t-1}\int_{\theta}V(\delta^*,\lambda\widetilde{\epsilon}(\theta))dF(\theta)$).

## 4.2 Solving for the Equilibrium

In this section we solve for the equilibrium given the patients' and the physicians' strategies described above. Although the equilibrium is achieved instantaneously, we adopt a sequential reasoning when solving for the equilibrium for presentation sake only.

### 4.2.1 Solving for the Equilibrium: The Simple Case

Recall that before competition begins, patients are equally allocated across physician types. Furthermore, suppose, for the time being, that switching costs $\kappa$ are arbitrarily small.

If the patient is currently with a physician for whom $\lambda^{\max}\widetilde{\epsilon}(\theta)<\epsilon^{\exp}(\theta)$, then the patient's threat of leaving for an outside physician is credible. This is because, in expectation, he can be made better off by seeking care from another physician. Given the distribution of $\lambda$s, a patient who leaves his current physician can expect to receive in the future at least $\epsilon^{\exp}(\theta)=\widehat{\lambda}\widetilde{\epsilon}(\theta)$, where $\widehat{\lambda}=\int_0^1\lambda d\Gamma(\lambda)$. This is because if he left he could expect to draw a physician of type $\widehat{\lambda}$ who would never be willing to provide less than $\widehat{\lambda}\widetilde{\epsilon}(\theta)$. As a result, all physicians of type $\lambda<\widehat{\lambda}$ will wish to provide the effort the patient could expect if he left for an outside physician, i.e., $\epsilon^{\exp}(\theta)=\widehat{\lambda}\widetilde{\epsilon}(\theta)$.

If the patient is currently with a physician for whom $\lambda^{\max}\widetilde{\epsilon}(\theta)\geq\epsilon^{\exp}(\theta)$, then he will not be willing to switch physicians because he can expect to draw a physician that could provide less than $\lambda^{\max}\widetilde{\epsilon}(\theta)$. Given that $\epsilon^{\exp}(\theta)=\widehat{\lambda}\widetilde{\epsilon}(\theta)$, all physicians of type $\lambda\geq\widehat{\lambda}$ should provide the effort

[12] Although we assume that the physician's future discounted expected utility associated with losing her patient is constant and exogenously given, we discuss the likely implications of this assumption in section 4.2.3.

determined by their ethical constraint (i.e., $\lambda\widetilde{\epsilon}(\theta)$) without risk of losing their patients.

Given the partial results provided above, we can see that effort levels should no longer be distributed between $[0, \widetilde{\epsilon}(\theta)]$ but rather between $[\epsilon^{\exp}(\theta), \widetilde{\epsilon}(\theta)]$=$[\widehat{\lambda\widetilde{\epsilon}}(\theta), \widetilde{\epsilon}(\theta)]$. If the $\lambda s$ are distributed according to a uniform distribution over [0,1], effort levels should thus be distributed between $\left[\frac{1}{2}\widetilde{\epsilon}(\theta), \widetilde{\epsilon}(\theta)\right]$ with half of the physicians treating with precisely $\frac{1}{2}\widetilde{\epsilon}(\theta)$. This is, however, not the full story. Suppose now that an individual has drawn a physician who treats him with exactly $\frac{1}{2}\widetilde{\epsilon}(\theta)$ (i.e., $\lambda^{\max} = \frac{1}{2}$). In such a case, the patient will have an incentive to leave because he can expect to receive at least $\epsilon^{\exp}(\theta) = \frac{1}{2}\left(\frac{1}{2}\widetilde{\epsilon}(\theta)\right) + \int_{\frac{1}{2}}^{1} \lambda\widetilde{\epsilon}(\theta)d\Gamma(\lambda) = \frac{5}{8}\widetilde{\epsilon}(\theta)$. Consequently, all physicians with a $\lambda < \frac{5}{8}$ should provide effort at precisely $\frac{5}{8}\widetilde{\epsilon}(\theta)$, while the rest should provide the effort determined by their ethical constraint. Thus, effort levels should now be distributed between $[\epsilon^{\exp}(\theta), \widetilde{\epsilon}(\theta)]$=$\left[\frac{5}{8}\widetilde{\epsilon}(\theta), \widetilde{\epsilon}(\theta)\right]$ with $\frac{5}{8}$ of the physicians providing $\frac{5}{8}\widetilde{\epsilon}(\theta)$. Using the same rationale, it can easily be shown that the only level of effort which survives in equilibrium is $\epsilon^*(\theta) = \epsilon^{\exp}(\theta) = \widetilde{\epsilon}(\theta)$, i.e., the equilibrium is characterized by a degenerate distribution of efforts. It is important to note that this rationale does not depend on the assumption of a uniform distribution of physician types.

Obviously, given that patients will always be provided with the desired effort ($\widetilde{\epsilon}(\theta)$), they will always choose the desired level of quantity ($\widetilde{q}(\theta)$). Thus, this equilibrium will be characterized by homogeneity in treatment and stable patient-physician relationships (i.e., patients will not move from one physician to another in equilibrium).

The above equilibrium, however, requires patient switching costs to be negligible and that no physician has any incentive to deviate and provide a level of effort below $\widetilde{\epsilon}(\theta)$, i.e. $\forall\lambda$,

$$V(\delta^*, \widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty} \beta^{t-1} \int_{\theta} V(\delta^*, \widetilde{\epsilon}(\theta))dF(\theta) \geq V(\delta^*, \lambda\widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty} \beta^{t-1} V_t^{DEV} \tag{17}$$

which is simply condition (16) where the expected effort level (i.e., the effort the patient could expect to receive if he left for an outside physician) is equal to the desired level of effort.

As before, given perfect competition in the insurance market, patients will choose the co-payment (and insurance premium) which maximizes expected utility. Furthermore, given that all

physicians will treat their patients with the desired levels of care ($\widetilde{\epsilon}(\theta)$) in equilibrium, physician expected utility will be identical across types (i.e., independent of their $\lambda$ parameter).

Proposition 1

*If switching costs are negligible and condition (17) is satisfied: (i) all physicians (irrespective of their $\lambda$) will treat their patients with the desired level of effort $\widetilde{\epsilon}(\theta)$; (ii) the patients will choose the desired level of medical care $\widetilde{q}(\theta)$; and (iii) patient-physician relationships will be stable.*

[INSERT FIGURE 1 HERE]

Because all physicians will provide the desired level of effort (i.e, for every $\theta$, $\epsilon^*(\theta) = \widetilde{\epsilon}(\theta)$), they will be compensated accordingly. Also, because the illness severity is not observable by the insurer, the equilibrium prospective payment $\delta^*$ must be illness independent and based on its expectation, i.e.,

$$\delta^* = \int_\theta \delta^*(\theta) dF(\theta) = \int_\theta c(\widetilde{\epsilon}(\theta)). \tag{18}$$

Notice that the prospective payment given in (18) is identical to that paid to physicians in the static framework. However, unlike the outcome in the static framework, the patients always receive desired levels of effort. Thus, no physician will receive a prospective payment which over-compensates her for effort provided (in expected terms). Furthermore, all physicians' ex post utility will be type independent.

Because the desired levels of quantity ($\widetilde{q}(\theta)$) and effort ($\widetilde{\epsilon}(\theta)$) will always be chosen in equilibrium, the actuarially-fair insurance premium is given by[13]:

$$\alpha^* = \pi \int_\theta ((1-\gamma^*) p\widetilde{q}(\theta)) dF(\theta) + \pi\delta^*. \tag{19}$$

Finally, the patient's ex post utility is given by:

$$U(I - \alpha^* - \gamma^* p\widetilde{q}(\theta), h(\theta, \widetilde{q}(\theta), \widetilde{\epsilon}(\theta))).$$

[13] In (19), $\gamma^*$ is the equilibrium co-payment, i.e., the one which balances the patient's expected utility gains of fuller insurance with the loss of a higher insurance premium.

In the above derivation, patients based their expectations about their current physician's future behaviour on $\lambda^{\max}$. We show in the Appendix that these are in fact the only expectations that survive in equilibrium under reasonable assumptions.

In the following two subsections, we examine the cases where: (i) switching costs are no longer trivial, and where (ii) condition (17) no longer holds for all physicians.

### 4.2.2 Solving for the Equilibrium: The Case where Patient Switching Costs are Non-Negligible

In the above section we began by showing that if a patient were currently with a physician identified by a $\lambda^{\max} < \widehat{\lambda}$, then he would be willing to leave for another physician if :

$$\sum_{t=2}^{\infty} \rho^{t-1} U^{Leave} - \sum_{t=2}^{\infty} \rho^{t-1} U^{Stay} > U(I-\alpha-\gamma pq^*, h(\theta, q^*, \epsilon)) - U(I-\alpha-\gamma pq^*-\kappa, h(\theta, q^*, \epsilon)). \quad (20)$$

Suppose now that the switching costs $\kappa$ are such that condition (20) exactly binds for a particular patient i.e., for this patient the present utility loss of switching from his current physician ($\lambda^{\max} < \widehat{\lambda}$) is just compensated by the expected future discounted utility gains of receiving the expected effort $\epsilon^{\exp}(\theta) = \widehat{\lambda}\widetilde{\epsilon}(\theta)$. Denote this particular patient's physician's $\lambda^{\max}$ as $\lambda^c(\kappa)$. All physicians with a $\lambda < \lambda^c(\kappa)$ should then behave like $\lambda^c(\kappa)$ in order to keep their patients. Consequently, a proportion $n$ of physicians (i.e., those with $\lambda < \lambda^c(\kappa)$) should provide effort such that their patients infer $\lambda^{\max} = \lambda^c(\kappa)$, while the rest should provide effort according to their own ethical constraint (i.e., $\lambda\widetilde{\epsilon}(\theta)$).[14] As a result, effort levels should be distributed between $[\lambda^c(\kappa)\widetilde{\epsilon}(\theta), \widetilde{\epsilon}(\theta)]$ with a proportion $n$ of physicians treating precisely at $\lambda^c(\kappa)\widetilde{\epsilon}(\theta)$. Under such an effort distribution, however, a patient with a physician who treats with $\lambda^c(\kappa)\widetilde{\epsilon}(\theta)$ could expect to receive $n\lambda^c(\kappa)\widetilde{\epsilon}(\theta) + \int_{\lambda^c(\kappa)}^{1} \lambda\widetilde{\epsilon}(\theta) d\Gamma(\lambda)$ if he left his current physician. Given this outside option, all physicians providing less than $n\lambda^c(\kappa)\widetilde{\epsilon}(\theta) + \int_{\lambda^c(\kappa)}^{1} \lambda\widetilde{\epsilon}(\theta) d\Gamma(\lambda)$ would, in the absence of switching costs, want to provide this amount to retain their patients. However, this is not the case in the presence of switching cost. That is, as before,

[14] As before, although it is possible for a physician characterized by a $\lambda^{\max} < \lambda^c(\kappa)$ to provide effort greater than the expected amount in the future (i.e., greater than $\lambda^c(\kappa)\widetilde{\epsilon}(\theta)$), those with a $\lambda > \lambda^c(\kappa)$ have no choice but to do so. Therefore, it is reasonable for the patient to base his expectations about his current physician's future treatments on his current physician's $\lambda^{\max}$.

there should exist a new critical effort level (i.e., a new critical physician type $\lambda^{c'}(\kappa)$) for which the patient is just indifferent between (i) staying with his current physician [i.e., receiving $\lambda^{c'}(\kappa)\widetilde{\epsilon}(\theta)$] and (ii) paying the switching cost and receiving expected effort [i.e., $n\lambda^{c}(\kappa)\widetilde{\epsilon}(\theta)+\int_{\lambda^{c}(\kappa)}^{1}\lambda\widetilde{\epsilon}(\theta)d\Gamma(\lambda)$]. Consequently, physicians with $\lambda$ smaller than $\lambda^{c'}(\kappa)$ should provide $\lambda^{c'}(\kappa)\widetilde{\epsilon}(\theta)$ in order to retain their patients while the rest should provide effort according to their own ethical constraint, i.e., $\lambda\widetilde{\epsilon}(\theta)$. As a result, effort levels should now be distributed between $[\lambda^{c'}(\kappa)\widetilde{\epsilon}(\theta),\widetilde{\epsilon}(\theta)]$. Using the same rationale, we can identify the equilibrium critical effort, say $\epsilon^{*}(\theta)$ (and its corresponding $\lambda^{*}(\kappa)$) which leaves a proportion of patients just indifferent between: (i) staying with their current physician and receiving $\epsilon^{*}(\theta)=\lambda^{*}(\kappa)\widetilde{\epsilon}(\theta)$; and (ii) paying $\kappa$, leaving and expecting to receive $n^{*}\lambda^{*}(\kappa)\widetilde{\epsilon}(\theta)+\int_{\lambda^{*}(\kappa)}^{1}\lambda\widetilde{\epsilon}(\theta)d\Gamma(\lambda)$. Thus, in equilibrium, a proportion $n^{*}$ of physicians (i.e., those characterized by a $\lambda<\lambda^{*}(\kappa)$) will provide $\lambda^{*}(\kappa)\widetilde{\epsilon}(\theta)$ while the rest (i.e., those characterized by $\lambda\geq\lambda^{*}(\kappa)$) will treat according to their ethical constraint $\lambda\widetilde{\epsilon}(\theta)$.

Proposition 2:

*In the presence of non-negligible switching costs, the equilibrium will be characterized by: (i) heterogenous effort levels with a proportion of physicians will treat beyond their ethical constraint while others treating according to their ethical constraint; and (ii) stable patient-physician relationships.*

[INSERT FIGURE 2 HERE]

Recall that, before competition begins, patients are equally distributed across physician types. Therefore, in the first period, patients have no information regarding their physician's type. However, they know that in equilibrium efforts will be distributed between $[\lambda^{*}(\kappa)\widetilde{\epsilon}(\theta),\widetilde{\epsilon}(\theta)]$ with the expected effort equal to:

$$n^{*}\lambda^{*}(\kappa)\widetilde{\epsilon}(\theta)+\int_{\lambda^{*}(\kappa)}^{1}\lambda\widetilde{\epsilon}(\theta)d\Gamma(\lambda). \tag{21}$$

Thus, given a particular illness severity $\theta$, the patient will choose the quantity of medical services $q^{*}$ based on this expected effort.

After one period, the patient's physician's $\lambda^{\max}$ is revealed. Given that the patient will remain

with the same physician for all periods and that this physician will provide effort equal to $\lambda^{\max}\widetilde{\epsilon}(\theta)$, the patient will choose $q^*$ based on $\lambda^{\max}\widetilde{\epsilon}(\theta)$ rather than (21).

It is also important to note that in order to ensure the participation of all physicians, the prospective payment ($\delta$) will need to compensate all physicians as if they were the most ethical type. That is, in equilibrium, switching costs will lead all but the most ethical physician to be over-compensated for the effort that they will provide. As a result, in the dynamic setting, the introduction of switching costs leads to reductions in effort without reductions in the prospective payment. The long-run actuarially-fair insurance premium $\alpha^*$ will be based on the equilibrium prospective payment and the expected medical expenditures.[15]

Although the effect of competition is dampened with the introduction of switching costs, competition nonetheless ensures a lower-bound on the effort provided ($\epsilon^*(\theta) = \lambda^*(\kappa)\widetilde{\epsilon}(\theta)$). Furthermore, as switching costs tend to zero, the proportion of physicians treating their patients with desired effort will tend to one. This may have important implications from a policy perspective. In fact, according to our model, any mechanism which reduces the costs (both psychic and financial) of moving from one physician to another will lead physicians to provide their patients with their desired levels of treatment.

It is important to mention here that, in the above, we assume that patients are risk-neutral with respect to their health. We make this assumption uniquely to keep things as simple as possible. It can be shown, however, that introducing risk-aversion in health is quite simple and leads to results which are qualitatively identical to those presented in this section (i.e., qualitatively identical to those found when introducing non-trivial switching costs).

#### 4.2.3 Solving for the Equilibrium: The Case Where Condition (17) No Longer Holds for all Physicians

The result in Proposition 1 where all physicians provide the desired level of effort $\widetilde{\epsilon}(\theta)$ relies not only on arbitrarily small switching costs but also on condition (17) not binding for all physicians. It

[15] Again recall that the equilibrium co-payment will be chosen by the patient to maximize expected utility.

is possible, however, that for some physicians, providing their minimal effort ($\lambda\widetilde{\epsilon}(\theta)$) thereby losing their patient yields greater expected utility than providing their patients with their expected effort and keeping them. By basing ourselves on (16) we can write down the following condition:

$$V(\delta^*, \lambda\widetilde{\epsilon}(\theta)) - V(\delta^*, \epsilon^{\exp}(\theta)) \geq \sum_{t=2}^{\infty} \beta^{t-1}(\int_{\theta} V(\delta^*, \epsilon^{\exp}(\theta))dF(\theta) - V_t^{DEV}) \qquad (17')$$

which simply states that the physician will (will not) deviate if the current-period benefits of deviating (i.e., providing minimal effort) are greater (smaller) than the expected discounted benefits of providing the expected effort. By examining the right-hand side of (17′), we can see two basic reasons why a physician may be willing to deviate. First, if the physician is relatively myopic (i.e., with a relatively small discount factor $\beta$), then the discounted expected benefits of keeping the patient will be too small to justify increased effort in the current period.[16] Second, if $\int_{\theta} V(\delta^*, \epsilon^{\exp}(\theta))dF(\theta) - V_t^{DEV}$ is relatively small (as would be the case if the physician were able to replace his patient relatively quickly because of, for example, excess demand) then a forward looking physician may be willing to deviate and lose his patient. We now turn our attention to the case where condition (17′) may in fact bind for some physicians.

In the absence of switching costs, let $\epsilon^{\lambda}(\theta)$ denote the maximum effort that physician $\lambda$ is willing to provide in order to keep her patient into the next period, i.e., for a physician $\lambda$:

$$V(\delta, \lambda\widetilde{\epsilon}(\theta)) + \sum_{t=2}^{\infty} \beta^{t-1} V_t^{DEV} = V(\delta, \epsilon^{\lambda}(\theta)) + \sum_{t=2}^{\infty} \beta^{t-1} \int_{\theta} V(\delta, \epsilon^{\lambda}(\theta))dF(\theta). \qquad (22)$$

Again assume that patients are initially equally distributed across physician types and that physicians are (for presentation sake) uniformly distributed across types. If the patient left his current physician, he could expect to draw a physician of type $E(\lambda) = \frac{1}{2}$ who would never be willing to provide less than $\frac{1}{2}\widetilde{\epsilon}(\theta)$. Given this, all physicians of type $\lambda < \frac{1}{2}$ and with $\epsilon^{\lambda}(\theta) < \frac{1}{2}\widetilde{\epsilon}(\theta)$ will wish to provide $\lambda\widetilde{\epsilon}(\theta)$ and lose their patients rather than provide the expected effort $\frac{1}{2}\widetilde{\epsilon}(\theta)$. We denote the group of physicians who are willing to provide their minimal effort even if this results

[16] If $\beta = 0$ (i.e., the physician does not care at all about the future) then the physician will always provide his minimal effort in equilibrium. That is, by setting $\beta = 0$, we obtain the same results as in the static framework of section 3).

in the loss of their patients as Group $A$. We further denote $\lambda^a$ as the physician type such that the physician is just indifferent between: (i) providing her minimal effort and losing her patient; and (ii) providing the expected effort and keeping her patient. Physicians with $\lambda < \frac{1}{2}$ but with $\epsilon^{\lambda}(\theta) > \frac{1}{2}\widetilde{\epsilon}(\theta)$ will wish to provide $\frac{1}{2}\widetilde{\epsilon}(\theta)$ in order to keep their patients. We denote the group of physicians who are willing to provide the expected effort in order to retain their patients as Group $B$. We further denote $\lambda^b$ to be the physician type such that the physician's minimal effort is equal to the expected effort. Obviously, all physicians with a $\lambda > \frac{1}{2}$ will wish to provide the effort determined by their ethical constraint without risk of losing their patients. We denote the group of physicians for whom the effort determined by their ethical constraint is greater than the expected effort as Group $C$. Thus the above describes a specific partition of physicians by type: Group $A$, Group $B$ and Group $C$.[17]

From the above description, we know that all patients whose current physician belongs to Group $A$ will leave. On the other hand, a patient whose current physician belongs to Group B (i.e., provides $\frac{1}{2}\widetilde{\epsilon}(\theta)$ at this stage) could expect to receive effort at least equal to:

$$\int_0^{\lambda^a} \lambda\widetilde{\epsilon}(\theta)d\Gamma(\lambda) + \frac{1}{2}\widetilde{\epsilon}(\theta)\int_{\lambda^a}^{\frac{1}{2}} d\Gamma(\lambda) + \int_{\frac{1}{2}}^{1} \lambda\widetilde{\epsilon}(\theta)d\Gamma(\lambda) \tag{23}$$

if he were to leave. Expected effort (23) is simply the expected effort provided by Group $A$, $B$ and $C$ weighted by the proportion of physicians in each group.

Suppose that a patient currently with a physician in Group $B$ is willing to leave his current physician because $\frac{1}{2}\widetilde{\epsilon}(\theta)$ is less than the expected effort provided in (23). Given the patient's outside option if he were to leave his current physician, a subset of physicians currently in Group $B$ will wish to increase their effort and provide the effort described in (23) (i.e., remain in Group $B$) while the others will wish to provide their minimal effort and lose their patients (i.e., join the newly formed Group $A$). Furthermore, a subset of physicians currently in Group $C$ whose minimal effort is less than (23) will wish to increase their effort to join the newly formed Group $B$. Thus, the

[17] For the case of a uniform distribution, the *size* of Group $A$ plus $B$ is equal to $\frac{1}{2}$, and the *size* of Group $C$ is equal to $\frac{1}{2}$.

patients' credible threat of leaving their current physician will lead to a new partition of Groups $A$, $B$ and $C$, where in general terms, a patient who is currently being treated by a physician in Group $B$ knows that if he left his current physician he could expect an effort level at least:

$$\int_0^{\lambda^a} \lambda\widetilde{\epsilon}(\theta)d\Gamma(\lambda) + \lambda^b\widetilde{\epsilon}(\theta)\int_{\lambda^a}^{\lambda^b} d\Gamma(\lambda) + \int_{\lambda^b}^{1} \lambda\widetilde{\epsilon}(\theta)d\Gamma(\lambda). \tag{24}$$

By building on the above logic, we can define the equilibrium where: (i) the patient's strategy is simply to remain with his current physician if the effort he receives is greater than or equal to the expected effort he would receive if he were to leave (given by (24)); and where (ii) the physician's strategy is simply to give the expected effort (or more if her ethical constraint binds) if the expected utility of providing expected effort is greater than providing her minimal effort and losing her patient.

Given the above strategies and derivation, the equilibrium will be characterized by a unique partition of physicians into Group $A$, Group $B$ and Group $C$, such that patients:

(i) who are treated with an effort level below $\lambda^b\widetilde{\epsilon}(\theta)$ (i.e., treated by a physician from Group $A$) will leave their current physician;

(ii) who are treated with an effort level greater than or equal to $\lambda^b\widetilde{\epsilon}(\theta)$ (i.e., with a physician from either Group $B$ or $C$) will remain with their current physician (i.e., $\lambda^b\widetilde{\epsilon}(\theta)$ is just equal to the expected effort in (24));

and where physicians:

(i) with $\lambda \leq \lambda^a$ will provide their minimal effort determined by their ethical constraint and lose their patients (these physicians may be thought of as revolving-door physicians);

(ii) with $\lambda^a < \lambda \leq \lambda^b$ will provide effort beyond that determined by their ethical constraint (i.e., will provide the expected effort given by (24)) and keep their patients;

(iii) with $\lambda > \lambda^b$ will provide their minimal effort determined by their ethical constraint and keep their patients.

Proposition 3:

*If Condition (17) binds for some physicians, the equilibrium will be characterized by three types of physicians: (i) physicians who treat their patients with their minimal effort (Group A) yet lose their patients; (ii) physicians who provide effort beyond the effort determined by their ethical constraint (Group B) and retain their patients; and (iii) physicians who provide effort levels determined by their ethical constraint (Group C) and retain their patients. Thus, heterogeneity in effort and some unstable physician-patient relationships may be sustained in equilibrium.*

[INSERT FIGURE 3 HERE]

Patients who have been randomly assigned a physician from either Group $B$ or $C$ in the first period, will choose the quantity of medical services $q^*$ based on the illness severity and the expected effort (because their physician's $\lambda^{\max}$ is not yet revealed). After the first period, however, they will be able to infer their physician's $\lambda^{\max}$ and make subsequent decisions based on $\lambda^{\max}\widetilde{\epsilon}(\theta)$.

Patients who have been randomly assigned a physician of Group $A$ in the first period, will also base their first-period's decision on the illness severity and the expected effort. Recall that a patient whose physician is of Group $A$ will leave for an outside physician. Thus, until the patient can infer that he is with a physician of either Group $B$ or $C$, he will continue to make his quantity decision in the same manner. Once the patient has been assigned a physician of Group $B$ or $C$ and has inferred his physician's $\lambda^{\max}$, he will base his quantity decision on his illness severity and on $\lambda^{\max}\widetilde{\epsilon}(\theta)$.[18]

Recall that a physician will deviate if the current-period benefits of deviating (i.e., providing minimal effort) are greater than the expected discounted benefits of providing the expected effort. The benefits of deviating may be relatively high in a situation with excess demand for physician services (i.e., in a case where physicians can readily replace their lost patients). By reducing the

[18] As in the previous cases, the equilibrium prospective payment $\delta^*$ will need to compensate for the effort provided by the most ethical physician in order to ensure the participation of all physicians. Furthermore, the equilibrium co-payment $\gamma^*$ will be chosen by the patient to maximize expected utility. Finally, the equilibrium actuarially-fair insurance premium $\alpha^*$ will reflect the expected medical expenditures.

future expected discounted utility associated with losing one's patient (which increases the expected benefits of maintaining one's patient), the equilibrium will tend to that described in Proposition 1. Thus, if condition (17) binds because of excess demand for physician services, increasing the supply of physicians may be a possible way to induce them to provide their patients with efforts which tend to their desired levels.

As stated in Proposition 3, in equilibrium, certain patients will be treated with effort which may be substantially below their desired level. These patients will leave for an outside physician. This equilibrium is, however, based on two implicit assumptions that we have made throughout the paper: (i) that a patient who leaves a physician is randomly assigned to another; and (ii) that the physician's future discounted expected utility of losing a patient is constant (i.e., is exogenous). Both of these conditions may be questionable over the long run. Although we do not address these issues in this paper, we next briefly discuss their likely implications.

As noted above, certain physicians (i.e., the aforementioned 'revolving-door' physicians) will always treat their patients with efforts below their desired levels and lose them. As a result, in every period, a certain percentage of the patients who leave Group $A$ will be randomly assigned to a new physician of either Group $B$ or $C$. Thus, over time, the discounted expected utility of losing a patient should increase, given that the pool of patients re-assigned to Group $A$ should decrease. This first effect may lead certain physicians of Group $A$ to provide higher effort in order to maintain their patients given that the future discounted expected utility associated with losing a patient has decreased (i.e., certain physicians of Group $A$ may wish to move to Group $B$). On the other hand, because a certain percentage of patients in Group $A$ will be randomly assigned to physicians of either group Group $B$ or Group $C$, these two latter groups will 'fill up' over time (i.e., physicians in these groups may no longer be able to accept new patients). Thus, over time, the likelihood that a patient who leaves a physician of Group $A$ will be reassigned to a physician also of Group $A$ should increase. This second effect should lead to an increase in the future discounted expected utility associated with losing a patient and thus lead to a decrease in effort. Endogenizing

the expected discounted utility of losing a patient in order to net out these two opposing effects is an interesting issue left for future research.

# 5 Conclusion

In this paper we examine the role of competition in the physicians market as a means of encouraging physicians to provide desired levels of care in a setting characterized by information asymmetry. In order to examine this role, we adopt a repeated game setting and solve for equilibria supported by credible threats. Our framework is distinguished, most notably, from the previous literature by this dynamic element as well as by introducing unobserved heterogeneity in the physicians market.

In the static framework, we show that all physicians will provide their minimum amount of unobservable effort, i.e., the amount determined by their ethical constraint. Consequently, the equilibrium is characterized by heterogeneity in effort (conditional on a given illness severity). In the dynamic framework, however, we show that competition may serve as an important mechanism to induce the desired provision of unobserved elements of medical care. More specifically, we show that under certain conditions competition may provide enough incentives for all physicians to provide their patients with their desired levels of care irrespective of the physician's ethical constraint. We also show that the introduction of switching costs may dampened the effect of competition yielding some heterogeneity in treatments. Competition, nonetheless, provides a lower bound on the provision of effort in the presence of such switching costs. Finally, we show that under certain conditions such as excess demand in the physicians market or myopic physicians, heterogeneity in treatments as well as some unstable patient-physician relationships may be supported in equilibrium.

This work may have several policy implications as conditions are provided for the provision of desired levels of non-observable (i.e., non-contractible) effort. By reducing switching costs (i.e., the psychic and monetary costs of moving from one physician to another) and/or by increasing the future discounted benefits of keeping one's patient (for example, by increasing the supply of

physicians), one may be able to support an equilibrium characterized by stable patient-physician relationships and the provision of desired levels of both observable and non-observable types of medical care.

It is worth noting that our results do not depend on the patient's observing their physician's effort prior to treatment decisions (as suggested by Ma and McGuire (1997)) nor does it require the patient's knowledge of their physician's type. These do, however, depend on the patient being able to perfectly infer his physician's effort ex-post. Relaxing this assumption, by, for example, introducing uncertainty in the link between illness severity, treatment and post-treatment health, is an other interesting extension left for future work.

## References

[1] Allard, M., Cremer, H., and M. Marchand (2001) 'Incentive Contracts and the Compensation of Health Care Providers,' *Economie Publique* 3, 37-54.

[2] Arrow, K (1963): 'Uncertainty and the Welfare Economics of Medical Care,' *American Economics Review* 53, 941-69.

[3] Blomqvist, Å (1991) 'The doctor as double agent: Information asymmetry, health insurance, and medical care,' *Journal of Health Economics* 10, 411-422.

[4] Danzon, P. (2000) 'Liability for Medical Malpractice,' in A.J. Culyer and J.P. Newhouse eds. Handbook of Health Economics (Amsterdam: Elsevier Science)

[5] Dranove, D. (1988) 'Demand inducement and the physician-patient relationship,' *Economic Inquiry* 26, 28-98.

[6] Ellis, Randall (1998) 'Creaming, Skimping and Dumping: Provider Competition on the Intensive and Extensive Margins,' *Journal of Health Economics* 17, 537-55.

[7] Ellis, Randall and Thomas McGuire (1986) 'Provider Behavior under Prospective Reimbursement: Cost Sharing and Supply' *Journal of Health Economics* 5, 129-51.

[8] Gal-Or, Esther (1999) 'Optimal Reimbursement and Malpractice Sharing Rules in Health Care Markets,' *Journal of Regulatory Economics* 16, 237-65.

[9] Gaynor, M. and W. Vogt (2000) 'Antitrust and Competition in Health Care Markets' in: A.J. Culyer and J.P. Newhouse, eds., *Handbook of Health Economics,* (Amsterdam: Elsevier Science North-Holland), Chapter 27.

[10] Léger, P.T. (2000) 'Quality Control Mechanisms under Capitation Payment for Medical Services,' *Canadian Journal of Economics* 33, 564-88.

[11] Ma, Ching-to Albert (1994) 'Health Care Payment Systems: Cost and Quality Incentives,' *Journal of Economics and Management Strategy* 3, 93-112.

[12] Ma, Ching-to Albert and Thomas G. McGuire (1997) 'Optimal Health Insurance and Provider Payment,' *American Economic Review* 87, 685-704.

[13] Rochaix, L., (1989) 'Information Asymmetry and Search in the Market for Physicians' Services,' *Journal of Health Economics* 8, 53-84.

[14] Wedig, Gerald, Mitchell, Janet B. and Jerry Cromwell (1989) 'Can Optimal Physician Behavior Be Obtained Using Price Controls?,' *Journal of Health Politics, Policy and Law* 14, 601-20.

# Appendix

Suppose that when forming expectations about his current physician's future effort, the patient does not use $\lambda^{\max}$ but rather uses the conditional expectation of $\lambda$ given $\lambda^{\max}$. That is, by inferring his physician's $\lambda^{\max}$, the patient knows that his physician's actual $\lambda \in [0, \lambda^{\max}]$ and therefore takes the expected value of his current physician's $\lambda$ based on this interval, i.e., $\lambda^1 = \int_0^{\lambda^{\max}} \lambda d\Gamma(\lambda)$. Thus,

equation (14) can be rewritten by replacing $\lambda^{\max}$ by $\lambda^1$. That is, the patient's strategy is simply to leave (stay with) his current physician if $\epsilon^{\exp}(\theta) > (\leq)\ \lambda^1\widetilde{\epsilon}(\theta)$.

We can now solve for the equilibrium using the same rationale as in section 4.2.

Given the patient's strategy, if the patient is currently with a physician for whom $\lambda^1\widetilde{\epsilon}(\theta) < \epsilon^{\exp}(\theta)$, then the patient's threat of leaving is credible. Furthermore, given the distribution of $\lambda$s, a patient who left his current physician could expect to receive in the future at least $\epsilon^{\exp}(\theta) = \widehat{\lambda}\widetilde{\epsilon}(\theta)$. This is because if he left he could expect to draw a physician of type $\widehat{\lambda}$ who would never be willing to provide less than $\widehat{\lambda}\widetilde{\epsilon}(\theta)$. As a result, all physicians of type $\lambda < 1$ will wish to provide the desired effort $\widetilde{\epsilon}(\theta)$. Being provided with effort $\widetilde{\epsilon}(\theta)$, the patient will infer $\lambda^{\max} = 1$ and $\lambda^1 = \widehat{\lambda}$. By doing so, the patient will not leave his current physician. If this is the case, however, then the patient's beliefs will never be confirmed. This is simply because by being provided with $\widetilde{\epsilon}(\theta)$, the patient's belief about his current physician's future efforts will be based on $\lambda^1\widetilde{\epsilon}(\theta)$ even though his current physician's actual future efforts will always be $\lambda^{\max}\widetilde{\epsilon}(\theta) = \widetilde{\epsilon}(\theta)$ ( i.e., the patient's beliefs are always incorrect). Given the physicians' actions, patients *should* base their current physician's future efforts on $\lambda^{\max}$ *rather than* $\lambda^1$. A similar proof can be derived for any other belief between $\lambda^1$ and $\lambda^{\max}$.

Now suppose that the patient bases his current physician's future effort on an effort level $\lambda^2 < \lambda^1 = \int_0^{\lambda^{\max}} \lambda d\Gamma(\lambda)$. By doing so, $\lambda^2$ will always be less than $\widehat{\lambda}$ and the patient will always leave his current physician for an outside one. If the patient always leaves his current physician, then no physician has any incentive to provide effort beyond that determined by their ethical constraint. However, because all patients leave their physicians in every period (i.e., all physician-patient relationships are unstable), then patients' expectations about their current physician's future behaviour are never *put to the test*. Even though patient's expectations are never 'disproved', and thus are not violated at equilibrium, they could easily be argued to be unreasonable. Furthermore, by having such expectations, patients would systematically leave physicians characterized by a $\lambda > \widehat{\lambda}$ (i.e., one who could never provide less than $\widehat{\lambda}\widetilde{\epsilon}(\theta)$) for the expected physician $\widehat{\lambda}$. Thus, in equilibrium,

patients would systematically leave more ethical physicians for less ethical ones. We exclude this possibility (i.e., these expectations) for these obvious reasons.

**FIGURE 1 :**

All physicians treat with $\tilde{\varepsilon}(\theta)$
Independent of $\lambda$

**FIGURE 2 :**

All physicians with $\lambda < \lambda^*(\kappa)$ treat with $\varepsilon^*(\theta) = \lambda^*(\kappa)\tilde{\varepsilon}(\theta)$

All physicians with $\lambda \geq \lambda^*(\kappa)$ treat with $\lambda\,\tilde{\varepsilon}(\theta)$

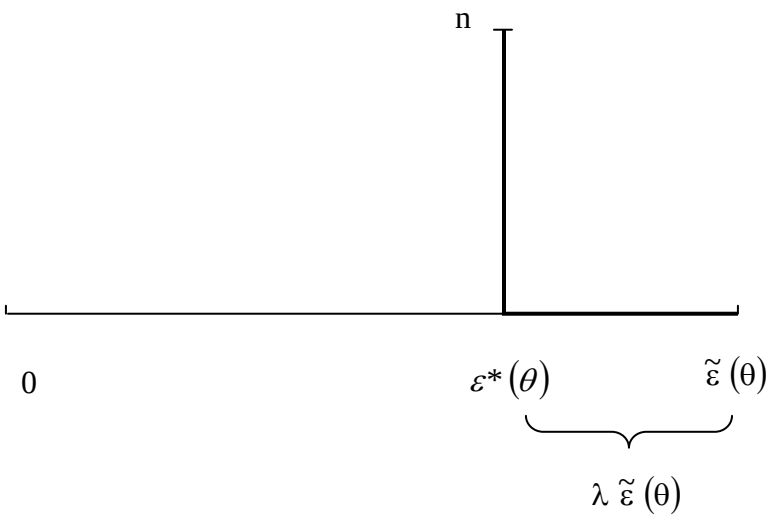
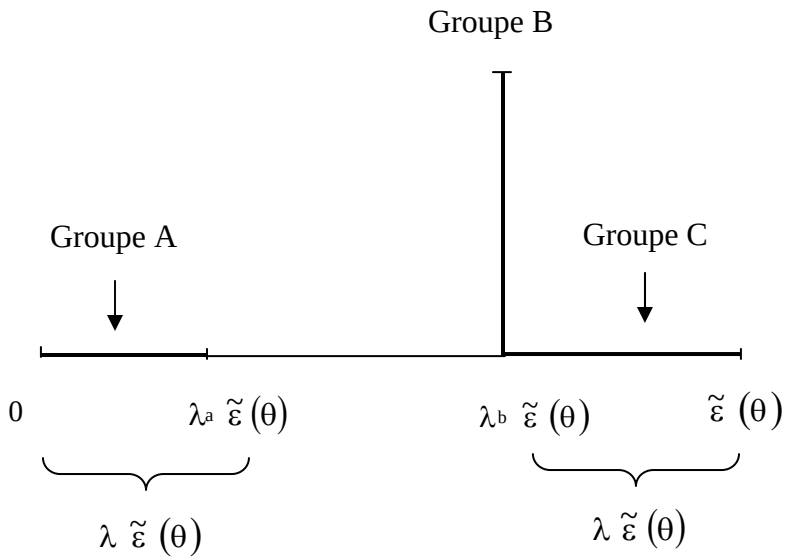


**FIGURE 3 :**


Groupe B
Groupe A
Groupe C
0
$\lambda_a \tilde{\varepsilon}(\theta)$
$\lambda_b \tilde{\varepsilon}(\theta)$
$\tilde{\varepsilon}(\theta)$
$\lambda \tilde{\varepsilon}(\theta)$
$\lambda \tilde{\varepsilon}(\theta)$

# Liste des cahiers de recherche publiés par les professeurs des H.E.C. 2003-2004

## Institut d'économie appliquée

IEA-03-01 GAGNÉ; ROBERT; LÉGER, PIERRE THOMAS. « Determinants of Physicians' Decisions to Specialize »,29 pages.

IEA-03-02 DOSTIE, BENOIT. « Controlling for Demand Side Factors and Job Matching: Maximum Likelihood Estimates of the Returns to Seniority Using Matched Employer-Employee Data »,24 pages.

IEA-03-03 LAPOINTE, ALAIN. « La performance de Montréal et l'économie du savoir: un changement de politique s'impose », 35 pages.

IEA-03-04 NORMANDIN, MICHEL; PHANEUF, LOUIS. « Monetary Policy Shocks: Testing Identification Conditions Under Time-Varying Conditional Volatility », 43 pages.

IEA-03-05 BOILEAU, MARTIN; NORMANDIN, MICHEL**.** « Dynamics of the Current Account and Interest Differentials », 38 pages.

IEA-03-06: NORMANDIN, MICHEL; ST-AMOUR, PASCAL. « Recursive Measures of Total Wealth and Portfolio Return », 10 pages.

IEA-03-07: DOSTIE, BENOIT; LÉGER, PIERRE THOMAS. « The Living Arrangement Dynamics of Sick, Elderly Individuals », 29 pages.

IEA-03-08: NORMANDIN, MICHEL. « Canadian and U.S. Financial Markets: Testing the International Integration Hypothesis under Time-Varying Conditional Volatility », 35 pages.

IEA-04-01: LEACH, ANDREW. « Integrated Assessment of Climate Change Using an OLG Model », 34 pages.

IEA-04-02: LEACH, ANDREW. « $_{Sub}$Game, set and match. Identifying Incentive Response in a Tournament », 39 pages.

IEA-04-03: LEACH, ANDREW. « The Climate Change Learning Curve », 27 pages.

IEA-04-04: DOSTIE, BENOIT; VENCATACHELLUM, DÉSIRÉ. « Compulsory and Voluntary Remittances: Evidence from Child Domestic Workers in Tunisia », 46 pages.

IEA-04-05: RENGIFO, E.W.; ROMBOUTS, J.V.K. « Dynamic Optimal Portfolio in a VaR Framework », 33 pages.

IEA-04-06: DOSTIE, BENOIT; TRÉPANIER, MATHIEU. « Return to Computer Use and Organizational Practices of the Firm », 41 pages.